

# VisualGPT Supplementary

Jun Chen<sup>1</sup>, Han Guo<sup>2</sup>, Kai Yi<sup>1</sup>, Boyang Li<sup>3</sup>, Mohamed Elhoseiny<sup>1</sup>

<sup>1</sup> King Abdullah University of Science and Technology,

<sup>2</sup>Carnegie Mellon University, <sup>3</sup>Nanyang Technological University

{jun.chen, kai.yi, mohamed.elhoseiny}@kaust.edu.sa

hanguo@cs.cmu.edu, boyang.li@ntu.edu.sg

## A. Additional implementation details

**Image and Word Features.** Following [1], we use a Faster R-CNN networks [10] with ResNet-101 [5] as a backbone to train on Visual Genome dataset [8], and we extract a 2048-dimensional feature vector for each object.

We use the Byte Pair Encoding (BPE) [12], which effectively incorporate sub-word information and is beneficial for dealing with out-of-vocabulary words. We employ learnable positional encoding and initialize token embedding from pretrained weights of GPT-2.

**Architecture and Hyperparameters.** We have 3 layers in the encoder and 12 layers in the decoder with 12 heads in each layer. The hidden size  $D$  in each layer is 768. We load the GPT-2 (small) pretrained weights, which has 117M parameters into the decoder. We use the learning rate of  $1e^{-4}$  under XE loss and  $1e^{-5}$  during the reinforcement learning. We train the models with the AdamW optimizer [9] and a batch size 25. The beam size is equal to 5. The threshold  $\tau$  is tuned on the validation set for different training data.

**Training Details.** We train all the models in two steps. We first train the models with cross-entropy (XE) loss and then finetune them using reinforcement learning. The cross-entropy loss  $\mathcal{L}_{XE}$  is the traditional autoregressive classification loss

$$\mathcal{L}_{XE} = - \sum_{t=1}^T \log((w_t | w_{1:t-1})) \quad (1)$$

where  $w_{1:T}$  represents the target ground truth sequence.

For reinforcement learning, we employ a variant of Self-Critical Sequence training [11]. Following [3], we sample  $L$  sentences,  $\hat{w}_{1:T}^1, \dots, \hat{w}_{1:T}^L$ , with beam search and use the mean reward from the  $L$  sentences as the baseline  $b$ . The gradient is

$$\nabla_{\theta} \mathcal{L}_{RL}(\theta) = -\frac{1}{k} \sum_{i=1}^L \left( (r(\hat{w}_{1:T}^i) - b) \nabla_{\theta} \log p(\hat{w}_{1:T}^i) \right) \quad (2)$$

where  $r(\cdot)$  represents the CIDEr-D reward.

Models	B-1	B-2	B-3	B-4
Direct Translation	26.5	11.6	4.5	1.9
ElJundi <i>et al.</i>	33.2	19.3	10.5	5.7
VisualGPT	<b>52.6</b>	<b>28.5</b>	<b>20.8</b>	<b>11.2</b>

Table 1. Arabic Image Captioning. Direct translation is to directly translate from English Caption to Arabic captions.

## B. Image Captioning in Low-resource languages Evaluation

Image captioning in low-resource languages suffers from having sufficient image-pairs to train a good-quality model. Currently, there are only very few major languages such as English or Chinese are well studied in image captioning domains, but a lot of low-resource languages have not been covered. Developing good multi-modal technologies for those low-resource languages opens considerable economic perspective and benefit a huge number of inhabitants in the world.

In this work, we attempt to evaluate our model on Arabic image captioning challenges, which is much less covered in the literature compared to English. There are very few good-quality image-caption pairs since it is very expensive to acquire the annotations. Some optional solutions are translating English captions to Arabic languages, but it requires to have a good language translation system and the translated captions need to maintain good grounding ability with the image contents, which is challenging to modern translation systems especially for those low-resource languages. We further evaluate our model on ElJundi *et al.*'s Arabic image captioning dataset [4] which is built based on Flickr8K [6] and contains 8K images. We follow their evaluation setting and train our VisualGPT on it. To adapt our VisualGPT on Arabic vocabulary, we instead use the pretrained GPT-2 in Arabic version [2].

The experimental results in Table 1. It shows that our VisualGPT can easily outperform the baseline models.

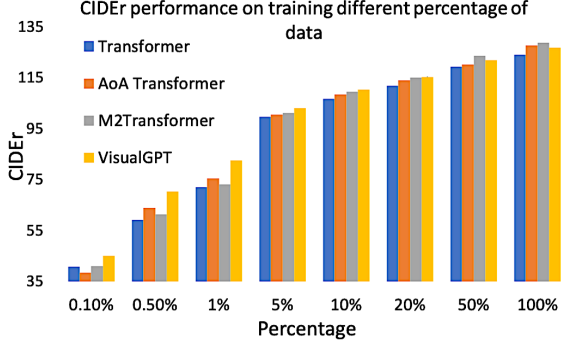


Figure 1. Evaluation on different percentage of COCO data

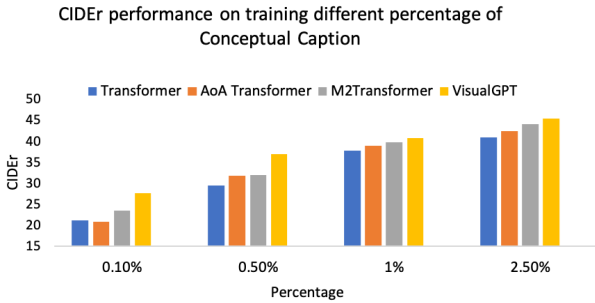


Figure 2. Evaluation on different percentage of Conceptual Captions

### C. Train VisualGPT with more COCO and Conceptual Caption Datasets

Figure 1 shows other results obtained by training networks on the 5%, 10%, 20%, 50% and 100% (82,783 images) MS COCO data. Figure 2 shows the performance with the data scaling up to 2.5% (82,958 images) Conceptual Captions, in which the dataset scale is similar to the whole COCO data. For MS COCO, VisualGPT outperforms other baseline models when we sample  $\leq 20\%$  training data. For Conceptual Caption, VisualGPT consistently outperforms all the baselines when we sample  $\leq 2.5\%$  training images. The whole experiments highlight our model’s effectiveness on low data regimes.

On the other hand, we should also notice that  $\mathcal{M}^2$  Transformer surpasses the VisualGPT’s performance when there are 50% and 100% COCO training data. But when we train with the same number of Conceptual images, VisualGPT continuously outperforms all the baselines. This leads us to think of the reason why VisualGPT show different performing behaviors on these two datasets. The difference between these two datasets is that the Conceptual Captions contain more diverse vocabularies and image contents. In

contrast, COCO captions only cover 80 common image objects. Therefore, the appearance frequency for each word in COCO is much higher than that in Conceptual Captions and COCO vocabulary diversity is also much lower than Conceptual Caption. We hypothesises the reason for this performance difference is that when the captions have a small coverage of each word, the caption generation will be benefited a lot from the GPT inherent knowledge and GPT can help the model quickly adapt into the new domain. But when there is a lot of in-domain data, the current image-captioning models can already generalize well on it and it potentially contradicts to the GPT original knowledge.

### D. Attention over Different types of words

We use the Spacy parser to detect the part-of-speech of words in captions and calculate the mean value of the visual attention score. The result is presented in Fig. 3. We found PoS that tend to visual content, like noun (0.71), verb (0.71) and adjective (0.72), have high visual attention scores, whereas linguistic PoS like pronoun (0.53), punctuation (0.58), and determiner (0.61) receive low attention.

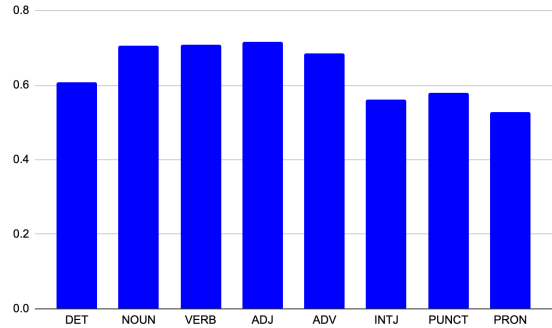


Figure 3. Attention Scores over different part-of-speech words

### E. More Qualitative Examples

In Figure 4, we provide more examples of visual attentions. Blue indicates high visual scores and red indicates low visual scores. We can observe that VisualGPT assigns higher scores to words like “steam engine”, “elephants”, “horse”, “lush” and “cabinets”, and it assigns low visual scores to determiners and prepositions like “to” and “at”.

We also show some examples of generated captions by our VisualGPT and several strong baseline models including Transformer (3 layers) [13],  $\mathcal{M}^2$  Transformer (3 layers) [3] and AoA Transformer [7] in the Table 2, Table 3 and Table 4. Overall, we can observe that our VisualGPT is able to describe the image content more accurately than the baseline models.






Image	Generated Captions	Ground Truth
	<p><b>Transformer:</b> a woman riding some skis on skis</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a couple of skiers are standing near the snow</p> <p><b>AoA Transformer:</b> a man with skis in the snow</p> <p><b>VisualGPT (ours):</b> a group of people walk on a snowy mountain</p>	<p><b>GT1:</b> the people are walking through snow in a wooded area</p> <p><b>GT2:</b> two people wearing skis traveling through the snow</p> <p><b>GT3:</b> a man is walking down a path covered in a snow</p> <p><b>GT4:</b> a couple is skiing through the snowy woods</p> <p><b>GT5:</b> a couple of people that are in a snowy field</p>
	<p><b>Transformer:</b> a street that has some street in it</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a traffic light over a street light under a traffic light</p> <p><b>AoA Transformer:</b> a street with people on a city street</p> <p><b>VisualGPT (ours):</b> a street with tall signs and traffic signs</p>	<p><b>GT1:</b> a yellow traffic light above a street next to houses</p> <p><b>GT2:</b> a street scene of an intersection with a street light</p> <p><b>GT3:</b> a stop light hanging over an intersection in a residential area</p> <p><b>GT4:</b> a traffic signal at an intersection is suspended on wire</p> <p><b>GT5:</b> a street intersection with a traffic light over it</p>
	<p><b>Transformer:</b> some pizza are sitting on a plate</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a plate with food and a knife on it</p> <p><b>AoA Transformer:</b> a plate of pizza on a table</p> <p><b>VisualGPT (ours):</b> a plate of bread are served on a table</p>	<p><b>GT1:</b> a batch of bread slices sitting on a plate</p> <p><b>GT2:</b> a plate with some pieces of bread on it</p> <p><b>GT3:</b> sliced french bread is on a plat that is lying on a table</p> <p><b>GT4:</b> bread that is sitting on a plate that is on a table</p> <p><b>GT5:</b> a white plate with lots topped with garlic bread</p>
	<p><b>Transformer:</b> two tennis player playing tennis on the ball</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a tennis player about to hit a ball</p> <p><b>AoA Transformer:</b> a baseball players on a game playing a game</p> <p><b>VisualGPT (ours):</b> a tennis player hits a ball with a racket</p>	<p><b>GT1:</b> a man holding a racquet on top of a tennis court</p> <p><b>GT2:</b> a man with a tennis racket reaches for a ball</p> <p><b>GT3:</b> a man with a tennis racket is running on a court</p> <p><b>GT4:</b> a young man is playing a game of tennis</p> <p><b>GT5:</b> a tennis player in a blue shirt runs toward a ball</p>
	<p><b>Transformer:</b> a group of birds that are standing in the grass</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a flock of birds perched in a tree branch</p> <p><b>AoA Transformer:</b> several giraffe are standing next to each trees</p> <p><b>VisualGPT (ours):</b> a bird standing in the middle of a pond</p>	<p><b>GT1:</b> a bird is perched a top a branch over a river</p> <p><b>GT2:</b> a bird sits on a branch above a stream</p> <p><b>GT3:</b> a bird on top of a tree branch over water</p> <p><b>GT4:</b> a picture of an outside region that appears incredible</p> <p><b>GT5:</b> a bird on a fallen branch in a body of water</p>

Table 2. Caption generated by our VisualGPT, Transformer,  $\mathcal{M}^2$  Transformer and AoA Transformer on 0.1% MS COCO data split

Image	Generated Captions	Ground Truth
	<p><b>Transformer:</b> several boats are sitting in the middle of a lake</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a boat filled with boats floating in the water</p> <p><b>AoA Transformer:</b> an empty boat that has water and water</p> <p><b>VisualGPT (ours):</b> a canal filled with boats in the water</p>	<p><b>GT1:</b> a blue boat docked on a green lush shore</p> <p><b>GT2:</b> a small marina with boats docked there</p> <p><b>GT3:</b> a group of boats sitting together with no one around</p> <p><b>GT4:</b> some boats parked in the water at a dock</p> <p><b>GT5:</b> boats sitting around the side of a lake by a tree</p>
	<p><b>Transformer:</b> pizza slices and pizza in a plate covered pizza</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> people sitting at a table eating pizza and other salad</p> <p><b>AoA Transformer:</b> two pizza eating a table with pizza on the table</p> <p><b>VisualGPT (ours):</b> a group of pizza on a iron plate with toppings</p>	<p><b>GT1:</b> a set of five pizzas sitting next to each other each with different toppings</p> <p><b>GT2:</b> a handful of prepared pizzas sit next to each other</p> <p><b>GT3:</b> five uncooked pizzas with a variety of different toppings</p> <p><b>GT4:</b> five unbaked pizzas that include various types of cheeses</p> <p><b>GT5:</b> five different pizzas are being prepared over a metal tray</p>
	<p><b>Transformer:</b> a dog holding a frisbee in the water</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a dog holding a frisbee in a body of water</p> <p><b>AoA Transformer:</b> a dog walking during a frisbee in a stone day</p> <p><b>VisualGPT (ours):</b> a dog walking through the water with a frisbee</p>	<p><b>GT1:</b> two dogs are playing on the beach catching a frisbee</p> <p><b>GT2:</b> of two dogs only one may be the victor</p> <p><b>GT3:</b> a dog catching a frisbee by another dog on a beach</p> <p><b>GT4:</b> dog jumping up in the air to catch a frisbee in the summer time</p> <p><b>GT5:</b> a dog jumping up into the air to catch a frisbee</p>
	<p><b>Transformer:</b> a group of people taking a child in a in a building</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a group of people in an airport with their hands</p> <p><b>AoA Transformer:</b> a picture of a young group of people standing for men</p> <p><b>VisualGPT (ours):</b> a group of people standing around a tv</p>	<p><b>GT1:</b> a group of men standing around a room</p> <p><b>GT2:</b> some people are waiting in a long room</p> <p><b>GT3:</b> people are standing in a room looking at a television screen</p> <p><b>GT4:</b> a person sitting on a bench while the rest look somewhere else</p> <p><b>GT5:</b> a man in red winter clothes sits on a bench with people behind him gather in front of a tv</p>
	<p><b>Transformer:</b> an elephant eating a elephant has a elephant</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> elephant with its trunk with their elephant with its trunk</p> <p><b>AoA Transformer:</b> two elephants standing at a lot of trees</p> <p><b>VisualGPT (ours):</b> three elephants standing next to some trees</p>	<p><b>GT1:</b> two adult elephants are surrounding a baby elephant</p> <p><b>GT2:</b> a baby elephant kneeling in front of two bigger elephants</p> <p><b>GT3:</b> a baby elephant and it's parents eat fruit</p> <p><b>GT4:</b> elephants eat fruit a baby elephant rummaging in the food</p> <p><b>GT5:</b> a pair of adult elephants with a baby elephant eat from a pile of fruit</p>

Table 3. Caption generated by our VisualGPT, Transformer,  $\mathcal{M}^2$  Transformer and AoA Transformer on 0.5% MS COCO data split



Image	Generated Captions	Ground Truth
	<p><b>Transformer:</b> a man in a suit and a woman standing in a shop</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a man is standing in a shop with a people holding people</p> <p><b>AoA Transformer:</b> a man is working on a bus in a</p> <p><b>VisualGPT (ours):</b> a group of people standing at an airport with their luggage</p>	<p><b>GT1:</b> several people are purchasing tickets at a bus station</p> <p><b>GT2:</b> some people are checking in at the ticket counter somewhere in asia</p> <p><b>GT3:</b> people waiting in line with luggage at a ticket counter</p> <p><b>GT4:</b> people are standing near an airport ticket kiosk</p> <p><b>GT5:</b> customers stand at a kiosk waiting for tickets</p>
	<p><b>Transformer:</b> a bus that is parked in front of a building</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a couple of people walking down the side of a street</p> <p><b>AoA Transformer:</b> a bus is parked in a city street</p> <p><b>VisualGPT (ours):</b> a white and blue bus is parked on the side of a city street</p>	<p><b>GT1:</b> people standing outside of a blue and white bus</p> <p><b>GT2:</b> an image of a tour bus that is picking people up</p> <p><b>GT3:</b> several people standing around buses and most wearing orange vests</p> <p><b>GT4:</b> a public transit bus pulling up to pick up passengers</p> <p><b>GT5:</b> a city bus at a stop waiting to pick up passengers</p>
	<p><b>Transformer:</b> a blue and white airplane flying through a sky</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> an air plane flying in the air</p> <p><b>AoA Transformer:</b> a plane airplane flying down in the sky</p> <p><b>VisualGPT (ours):</b> a plane is flying in the air over the trees</p>	<p><b>GT1:</b> there 's and airplane in the sky flying over some trees</p> <p><b>GT2:</b> a large plane is flying over a crowd of trees</p> <p><b>GT3:</b> a aeroplane soaring high in the sky above the trees</p> <p><b>GT4:</b> a passenger plane flies in the sky over a forest</p> <p><b>GT5:</b> an airplane is seen flying over several trees</p>
	<p><b>Transformer:</b> a white toilet sitting in a white bathroom next to a sink</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a cat sitting in the toilet</p> <p><b>AoA Transformer:</b> a bathroom with a toilet and a sink</p> <p><b>VisualGPT (ours):</b> a cat sitting on top of a bathroom sink</p>	<p><b>GT1:</b> a cat climbing into a bathroom sink looking at someone</p> <p><b>GT2:</b> a cat looks up as it stands in the bathroom sink</p> <p><b>GT3:</b> a large cat stands inside of a clean bathroom sink</p> <p><b>GT4:</b> cat is caught stepping in to the bathroom sink</p> <p><b>GT5:</b> a cute kitty cat in the sink of a bathroom near a brush and other items</p>
	<p><b>Transformer:</b> a little girl is eating a birthday cake</p> <p><math>\mathcal{M}^2</math> <b>Transformer:</b> a child and a child are sitting at a table with table with table</p> <p><b>AoA Transformer:</b> two children sitting at a table with a laptop computer</p> <p><b>VisualGPT (ours):</b> a woman and a girl sitting at a table with a birthday cake</p>	<p><b>GT1:</b> a woman and child stand next to a table with cake on it</p> <p><b>GT2:</b> a lady standing near the table with a baby is posing for the camera</p> <p><b>GT3:</b> a woman stands beside a baby in a high chair a table is set with a birthday cake and champagne</p> <p><b>GT4:</b> a woman setting up her house for a party</p> <p><b>GT5:</b> a person standing next to a child in a booster seat</p>

Table 4. Caption generated by our VisualGPT, Transformer,  $\mathcal{M}^2$  Transformer and AoA Transformer on 1% MS COCO data split

	GT: the large red flower is inside of a clear glass vase										
Ours	a	red	vase	of	roses	sitting	on	top	of	a	glass
attention	0.8	0.93	0.94	0.64	0.87	0.84	0.67	0.55	0.57	0.43	0.86
	GT: a tennis player jumps and hits a ball										
Ours	a	tennis player	jumping	on	a	tennis	court	holding	a	ball	
attention	0.7	0.77	0.75	0.72	0.67	0.64	0.89	0.79	0.74	0.6	0.76
	GT: a motorcycle parked next to a white building										
Ours	a	motorcycle	parked	next	to	a	building				
attention	0.6	0.78	0.85	0.74	0.34	0.6	0.75				
	GT: a small boats in a body of water										
Ours	a	large	boat	sits	on	a	field	with	a	lake	
attention	0.6	0.77	0.78	0.83	0.71	0.6	0.74	0.66	0.63	0.73	
	GT: a kitchen with wooden cabinets a sink and a dish washer										
Ours	a	kitchen	with	a	white	cabinets	and	a	sink		
attention	0.73	0.86	0.8	0.7	0.9	0.91	0.8	0.8	0.9		
	GT: a train sitting under a display inside a building										
Ours	a	steam	engine	sitting	in	a	display				
attention	0.69	0.84	0.79	0.8	0.7	0.6	0.83				
	GT: two captive elephants stand bored behind the fake stone fence										
Ours	elephants	standing	next	to	a	stone	fence				
attention	0.8	0.74	0.77	0.47	0.5	0.77	0.76				
	GT: a white horse standing in a field on top of grass										
Ours	a	white	horse	grazing	on	a	lush	green	field		
attention	0.67	0.75	0.83	0.74	0.65	0.66	0.85	0.8	0.77		
	GT: a man in a restaurant smiling while holding up a camera										
Ours	a	man	in	a	store	looking	at	his	camera		
attention	0.65	0.69	0.72	0.67	0.77	0.65	0.47	0.49	0.7		
	GT: a man sitting on a bench next to a few bags										
Ours	a	young	man	holding	a	backpack	on	a	bench		
attention	0.7	0.82	0.74	0.7	0.54	0.84	0.59	0.55	0.83		

Figure 4. More examples of visual attention for each word in generated captions. High visual scores are in blue and low scores in red.

## References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1
- [2] Wissam Antoun, Fady Baly, and Hazem Hajj. Aragpt2: Pre-trained transformer for arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, 2021. 1
- [3] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 1, 2
- [4] Obeida ElJundi., Mohamad Dhaybi., Kotaiba Mokadam., Hazem Hajj., and Daniel Asmar. Resources and end-to-end neural network models for arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, pages 233–241. INSTICC, SciTePress, 2020. 1
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1
- [6] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1
- [7] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 2
- [8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1), 2017. 1
- [9] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 1
- [10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 1
- [11] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1
- [12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*. The Association for Computer Linguistics, 2016. 1
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2